International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V73I3P106

Volume 73 Issue 3, 73-94, March 2025
© 2025 Seventh Sense Research Group®

Review Article

A Literature Review of Machine Learning Techniques for Cyberbullying Detection in Arabic Social Media Text

Bader Azi Alanazi[1], Chin-Teng Lin[2]

[1,2]University of Technology Sydney (Australia).
[1]Jouf University (Kingdom of Saudi Arabia).

[1]Corresponding Author : baenzi@ju.edu.sa

Received: 7 May 2024 Revised: 01 October 2024 Accepted: 13 February 2025 Published: 28 March 2025

Abstract *- Social media, including platforms such as X (known as Twitter), have become an integral part of our daily lives, serving as a primary means of communication for people globally. Social media are widely used to share thoughts, viewpoints, and critiques. However, the accessibility of these platforms can sometimes lead to misuse, paving the way for cyberbullying–a damaging form of online harassment. Although much research has been conducted on cyberbullying detection in the English language, there is a noticeable research gap regarding the Arabic language. Spotting cyberbullying in Arabic posts on X (Twitter) can help make the platform safer and friendlier for Arabic users while highlighting the harm it inflicts. Cyberbullying instances can be identified and categorized using tools such as Natural Language Processing (NLP) and machine learning algorithms. This paper reviews studies that leverage Machine Learning to identify instances of cyberbullying in Arabic.*

Keywords *- Machine Learning, Cyberbullying, Arabic text, Twitter, Natural language processing.*

1. Introduction

Social media networks, particularly platforms like Twitter and Facebook, have seen exponential growth in recent years. In Saudi Arabia alone, Twitter boasts 18.33 million users, making it one of the most widely used platforms in the region [1]. With over 330 million monthly users globally, Twitter has become a space for individuals to express their views, share insights, and engage in discussions [2]. However, the rise in social media usage has also brought about significant negative consequences, most notably the surge in cyberbullying. Cyberbullying refers to the use of digital platforms-such as social media, emails, and instant messages-to harass, intimidate, or demean others. This issue has reached critical levels, necessitating immediate intervention. The impact of cyberbullying is particularly pronounced among young users, with severe consequences such as anxiety, depression, and suicidal tendencies often reported [3]. On Twitter, cyberbullying takes on various forms, including the spread of harmful rumors, public shaming, and the posting of derogatory content [4]. Social media platforms' anonymity further exacerbates this issue, as it emboldens individuals to engage in harmful behaviors without fear of repercussion. Moreover, cyberbullying on platforms like Twitter is not only a personal concern but also a societal one, as it often reflects and amplifies broader issues of discrimination, prejudice, and toxicity within online communities [5]. Recent advancements in machine learning, a branch of artificial intelligence, have shown promise in detecting cyberbullying by automatically identifying harmful patterns in online text [6]. These models, trained to recognize offensive language and threatening behavior, offer a proactive solution to addressing cyberbullying [7]. Most of the work, however, has focused on detecting cyberbullying in English. Research targeting Arabic-language cyberbullying, particularly on Twitter, remains limited [8].The majority of existing studies on Arabic cyberbullying detection have employed traditional machine learning techniques such as Support Vector Machines (SVM), and Naive Bayes (NB) classifiers [8-12], achieving accuracy rates as high as 95.9%.

However, Existing studies have not detected the severity of cyberbullying. This critical gap becomes even more pronounced when considering Arabic's cultural and linguistic nuances, particularly in the Saudi dialect. In addition, these studies often grapple with imbalanced datasets, resulting in biased outcomes in identifying the full extent of cyberbullying incidents. Moreover, cultural and linguistic differences in Arabic further complicate cyberbullying detection. For example, certain expressions deemed offensive in Arabic cultures may not have the same connotations in other languages. Words like "كلب" (dog) or "حمار" (donkey), while considered highly insulting in Arabic, may not have the same impact in other linguistic contexts [11]. These cultural variations, coupled with the inherent difficulties of Arabic language processing, highlight the critical gap in the current research, particularly when it comes to detecting the severity

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

of cyberbullying in the Saudi dialect. This review addresses these challenges by exploring the current state of research on Arabic cyberbullying detection, identifying the existing gaps, and proposing future directions. Specifically, it focuses on the need for a robust system capable of detecting the severity of cyberbullying in the Arabic-speaking world, focusing on Twitter in the Saudi Dialect, and overcoming the challenges posed by imbalanced datasets. The structure of this paper is as follows: Section 2 provides the necessary background information. Section 3 explains the methodology and approach used in this study. Section 4 explores the growing concern of cyberbullying on social media. Section 5 offers a comprehensive literature review, highlighting gaps in existing research. Section 6 examines the distinctions between Arabic and English cyberbullying detection models. Section 7 addresses the limitations and challenges faced by current approaches. Finally, Section 8 outlines the problem statement, while Section 9 concludes with recommendations for future research.

2. Background

2.1. Machine Learning Text Analysis

Machine Learning (ML), a branch of artificial intelligence, involves designing algorithms that learn from experience to improve specific tasks. These algorithms can study data spot patterns and make predictions or judgements without explicit programming instructions [13]. This unique capability makes ML very useful for various tasks, such as filtering spam, identifying images, diagnosing medical conditions, and analyzing stock market trends.

In recent years, ML models have gained the ability to perform various text classification tasks. Some well-known models include the Naive Bayes and Support Vector Machines. Furthermore, more advanced deep learning approaches, such as convolutional neural networks and recurrent neural networks, have demonstrated their ability to classify texts [14]. These models were built on datasets filled with text documents, each labelled with multiple types or categories. The goal of the learning phase is to identify and understand the training data patterns associated with specific labels [15]. Therefore, where does this start? It starts with feature extraction, identifying and pulling out key traits or "features" from the text data. This step is a game changer because the type and quality of the features can significantly influence the effectiveness of the machine learning models. Once the features were extracted, the models used the processed data for the training. This is where they learn the link between the features and labels. Next, the models were tested to determine how accurately they could predict the correct labels for a new dataset, also known as the testing or validation set [16].

2.2. Machine Learning Techniques

As Internet and social media use continues to rise, cyberbullying detection has climbed the priority ladder [17, 18]. Owing to its ability to automate the identification of abusive content, machine learning is shining a ray of hope in the battle against this pervasive issue. Several techniques under the machine learning umbrella have been employed to tackle this challenge, ranging from supervised to unsupervised and deep learning methods.

2.2.1. Supervised Learning Techniques

Supervised learning, a typical machine learning approach, involves training a model using a dataset with labelled instances, empowering it to make predictions or judgments without human guidance [19]. When pinpointing cyberbullying incidents, these techniques rely on datasets in which human reviewers have already marked instances of cyberbullying. Algorithms, such as Naive Bayes, Support Vector Machines, and Decision Trees, are among the major names in this space [20]. Their technique lies in spotting patterns in the labelled data, which they then use to decide if new data fits the "bullying" or "non-bullying" category.

2.2.2. Unsupervised Learning Techniques

Unsupervised learning approaches do not require pre-labelled data, which makes them useful when labelling data is challenging or unfeasible [19]. When spotting cyberbullying, these techniques aim to uncover hidden patterns or structures within the data that hint at bullying behaviour [21]. For example, clustering algorithms such as K-means can combine similar text messages, potentially revealing clusters of content linked to bullying.

2.2.3. Semi-Supervised Learning Techniques

The learning approach, known as semi-supervised learning, is a combination of supervised and unsupervised learning. This is useful when detecting instances of cyberbullying, and the labelled data is difficult to obtain. This technique is particularly beneficial when more tagged data are required [22]. Semi-supervised learning can boost the performance of a model using a limited amount of labelled data and a more significant chunk of unlabeled data. Recent studies have shown that models using semi-supervised learning offer promising results in sniffing cyberbullying incidents [23].

2.2.4. Deep Learning Techniques

Deep learning is a type of machine learning that uses Artificial Neural Networks to sift data. Owing to their numerous hidden layers, these networks can spot intricate patterns hidden in data [24]. Deep learning methods can effectively identify instances of cyberbullying [25, 26]. Various deep learning techniques[27-29], such as Long Short-Term Memory networks (LSTMs) and Convolutional Neural Networks (CNNs), have been used to detect cyberbullying. These techniques capture the context and semantic essence of the textual content. These models oversee vast volumes of high-dimensional data, such as text from social media platforms. They can effectively manage and organize data.

2.3. Natural Language Processing

Natural Language Processing (NLP) is crucial for text analysis. It represents the intersection of linguistics and artificial intelligence, concentrating on the intriguing interactions between human language and computers [30]. NLP's mission is to dissect and interpret human language meaningfully and helpfully [31]. This is about training machines in the art of human conversations. This conversation can be written or spoken; NLP is essential to decode it. This field connects human communication with digital data processing, enabling computers to grasp, interpret, and reproduce human language. NLP requires a profound understanding of the subtleties embedded in human language. This encompasses understanding syntax, the organization of words in a sentence, and semantics, which is the meaning extracted from sentences and the context surrounding them [32]. Grappling with the complexities and ambiguities of human language to effectively understand and process it can be challenging but necessary [33].

In addition, NLP techniques have automated various tasks that require manual labour [30, 32]. This includes language translation, speech recognition, sentiment analysis, and data extraction [34]. With the ever-growing influx of unstructured text input, the need for NLP to convert it into structured, helpful information has become more precise. This underscores the crucial role of NLP across various sectors, including healthcare, banking, and customer service, where decisions often hinge on textual input.

2.4. The Arabic Language and NLP

2.4.1. Characteristics and Challenges of the Arabic Language

Arabic is among the six languages officially recognized by the United Nations. This Semitic language, known for its intricate morphology and rich syntactic structure, is spoken by 400 million individuals across 22 nations [35]. By 2013, approximately 135 million people had utilized the Internet in Arabic [36].

The Arabic writing system comprises 28 letters written from right to left. The language features numerous dialects, a wide-ranging morphology capable of addressing various subjects, and the use of diacritical marks [37-40]. Arabic has a complex structure, and its vocabulary is mainly based on a group of roots consisting of three, four, or five letters [36]. Three-letter roots were the most frequently used. Arabic has three main classes of words: nouns, which include adjectives and adverbs; verbs; and particles. In official Arabic writing, sentences are typically demarcated by commas and periods. These factors contribute to language's complicated nature. Arabic's complex system of root words and patterns adds another layer of difficulty to language processing.

According to [36, 41], Arabic comes in two primary forms: Standard and Dialectal. The former is subdivided into Classical Arabic (CA) and Modern Standard Arabic (MSA), whereas Dialectal Arabic encapsulates all variants spoken in everyday life across different countries. Each dialect strays slightly from the Standard Arabic [42]. Modern Standard Arabic, in particular, is the norm for written and formal communication and is typically found in books, newspapers, news broadcasts, formal speeches, and movie subtitles, continuing similarly in Arabic-speaking countries.

It is important to note that substantial grammatical and vocabulary differences between MSA and the numerous colloquial dialects further complicate the application of NLP methods. These dialects can be distinguished from MSA based on their proximity to one another. Discrepancies in spelling, particularly in user-generated online content, create extra hurdles that machine-learning models need to overcome [42].

3. The State of Arabic NLP

Arabic, with its intricate structure and extensive vocabulary, poses many challenges to NLP. However, despite the inherent complexities of the right-to-left script, many dialects, and prevalent diacritics, progress has been made in Arabic NLP over the past few years. Many tools and resources are now available, which empower us to tackle NLP tasks involving Arabic text [8]. These tools and resources include Arabic language corpora, stemming algorithms, and part-of-speech taggers [43]. These leaps have opened doors to various NLP tasks in Arabic, including sentiment analysis, named entity recognition, speech recognition, and text categorization. These efforts have considerably enhanced our capacity to process and analyze Arabic text, which has dramatically improved [44].

4. Cyberbullying

4.1. The Definition of Cyberbullying

According to [45], bullying in its digital version can be referred to as cyberbullying, which can happen on a variety of digital channels, such as social media, email, and instant messaging. It is a deliberate act of aggression that either a person or a group commits against victims who are unable to defend themselves. Adolescents and young individuals are often the primary victims of cyberbullying due to their greater susceptibility to current technologies, such as social media platforms. One form of abuse, "cyberbullying", involves one person being rude to another online to the point where the target is insulted or offended [46]. Cyberbullying is defined as the deliberate, persistent, and hostile use of technology to hurt or harass others [3]. According to [47-50] cyberbullying is defined as deliberately damaging the conduct of a group or individual over time using modern digital technologies to attack victims who are powerless to protect themselves. Cyberbullying varies from traditional bullying in that it occurs every day of the week for 24 hours a day. It is perpetrated through various means, including the transmission of text messages, dissemination of gossip, posting of online content, and distribution of humiliating images and footage via social networking platforms [51]. Cyberbullying has a significant adverse effect on victims' mental health and wellbeing and can even lead to suicidal ideation and depression in some cases. As social media platforms gain popularity, the number of reported cases of cyberbullying is also growing. As a result, it is more important than ever to discover efficient methods to identify and stop destructive behaviour known as cyberbullying.

5. The Types of Cyberbullying

According to [3, 52-55] various forms of cyberbullying are prevalent on social media platforms, including Twitter. Cyberbullying is the act of harassing or intimidating others online, which commonly takes place on popular social media platforms, including email, Facebook, Instagram, Twitter, blogs, and YouTube, where accessing the internet is easy. The research results identified the following categories of cyberbullying:

- Harassment: Harassment involves regularly sending offensive, rude, or threatening texts to the victim. Examples include insults, name calls, or inappropriate remarks about the victim's appearance, religion, or other personal characteristics.
- Denigration: Denigration is the spread of false or malicious rumours about victims harming their images. Examples include posting embarrassing or offensive photos, videos, or remarks about the victim and creating fake profiles to impersonate and mock the victim.
- Impersonation: Impersonation occurs when the bully impersonates the victim by hacking their account or establishing a fake profile with their personal information. While pretending to be a victim, the bully can send offensive or harmful messages, share inappropriate content, or participate in malicious activities.
- Exclusion: Exclusion is the intentional act of excluding the victim from online organizations or activities, such as group chats or events, resulting in feelings of isolation and rejection.
- Cyberstalking: Cyberstalking is the repeated, unwanted monitoring and tracking of a victim's online actions, frequently followed by harassment or threats. Obtaining personal information, such as the victim's home location or phone number, and using it to intimidate or harass the victim are examples.
- Trickery: Trickery occurs when someone deceives you to trust them to share secrets or confidential private information, which they will then use to share with the public online.
- Fraping: When someone pretends to be the proprietor of your social media account and posts inappropriate content to trick others into believing that they posted the content.
- Dissing: When someone publishes details or conjectures about another individual with the purpose of harming their popularity or reputation.

Detecting and preventing these forms of cyberbullying on social media platforms such as Twitter is crucial for protecting the well-being of users and promoting a positive online environment.

6. Impact of Cyberbullying

According to [2], cyberbullying is a widespread problem that can seriously affect the self-esteem and mental health of people who experience it. Cyberbullying, a type of online harassment, can leave victims feeling lonely, defenceless, and helpless, thus leaving lasting emotional scars. The harmful effects of cyberbullying can emerge in various ways, including increased anxiety, sadness, and suicidal thoughts or actions. It is critical to recognize the seriousness of cyberbullying and work towards appropriate solutions to protect individuals from its negative consequences (Figure 1).

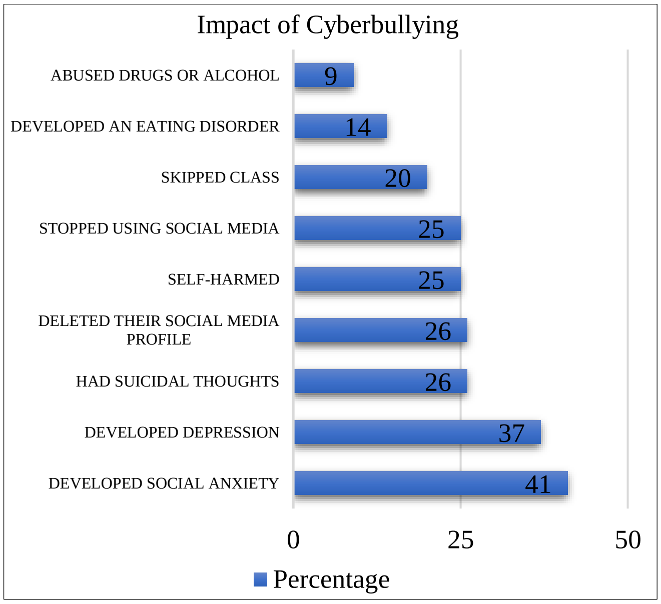

Fig. 1 The impacts of cyberbullying [8]

Cyberbullying can have both long- and short-term impacts on victims. One of the most severe consequences is mental health; cyberbullying can lead to anxiety, depression, and other mental health problems. Victims may feel isolated, which aggravates their symptoms. Cyberbullying is associated with suicide in some situations, making it a severe public health risk [56].

Cyberbullying can also affect academic achievement because victims may find it challenging to focus on and miss school or other crucial activities [57]. Cyberbullying can harm a victim's reputation both online and in real life [58]. When inaccurate or harmful information is shared online, it can be difficult to remove or amend, perhaps resulting in long-term ramifications. Victims may also feel socially isolated because they hesitate to engage in online or in-person social contact [59].

Not only the victims of cyberbullying are affected, but also their families, the community, and friends. Those who witness such acts may feel powerless because of their inability to step in, leading to substantial emotional and mental health effects. Moreover, cyberbullying can cultivate a toxic online environment in which individuals feel emboldened to indulge in harmful conduct without worrying about its potential consequences. Cyberbullying can stifle open communication and create a less welcoming and inclusive atmosphere in online spaces. As shown in Figure 2, this study demonstrates the fluctuating rates of cyberbullying victimization examined over the years [58]. Approximately 31% of the students who participated in the 13 most recent research reported experiencing cyberbullying at least once in their lifetime, as shown in this study. The incidence of cyberbullying perpetration has also differed among the research studies they have undertaken. Another study [60] they focused on two subsets of US school students: those regularly absent from school and those who chose home-schooling programs. Their research aimed to determine whether cyberbullying resulted in children missing school or, in more severe instances, quitting school. Their study found that nearly 17.5% of the students frequently stayed away from school because of adverse effects. Cyberbullying emerged as the fifth most common reason parents chose to home-school their children. These findings shed light on the substantial issue of cyberbullying, which affects many individuals.

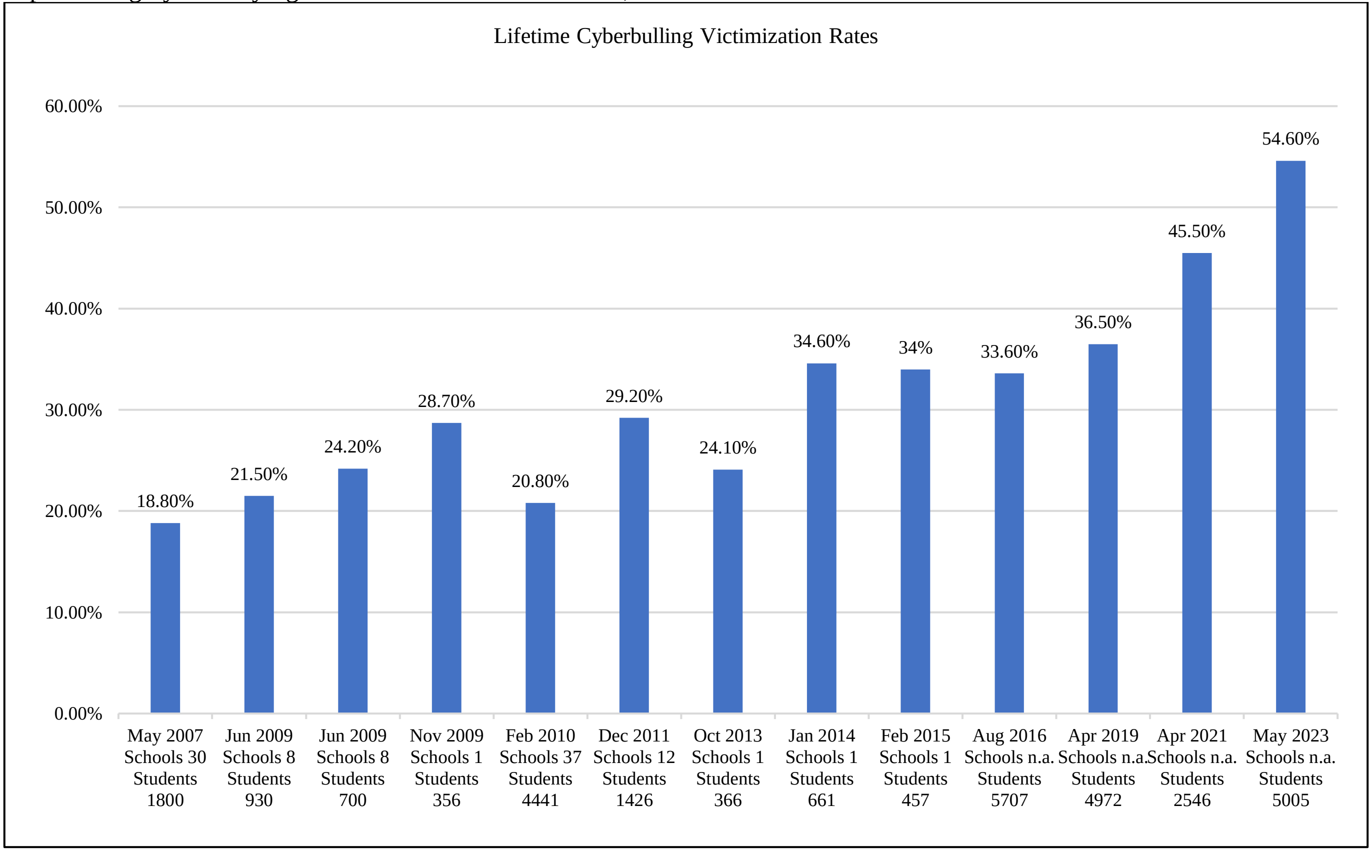

Fig. 2 Cyberbullying victimization [59]

In 2013, the National Family Safety Program in Saudi Arabia (overseen by the Ministry of Labor and Social Development) conducted a survey of a sample of 15264 high school students, revealing that 25% of participants had experienced cyberbullying [8]. Consequently, Saudi Arabia has recognized the damaging consequences of cyberbullying.

In response, the National Family Safety Program initiated a campaign in 2014 called the National Project for Cyberbullying Control, aiming to mitigate these effects and raise public awareness of cyberbullying to help students build their self-esteem.

7. Literature Review Method

To conduct this comprehensive literature review, we systematically searched for research articles using specific keywords relevant to our study. The keywords included “cyberbullying,” “cyberbullying detection,” “social media,” “machine learning,” “deep learning,” and “Arabic tweets.” These terms were used to query several established scientific databases: IEEE Xplore, ACM Digital Library, Scopus, SpringerLink, and Google Scholar. Clearly defined criteria guided the selection of articles, following systematic review principles outlined by [61]. Figure 3 shows the steps of the review and the number of included and excluded articles.

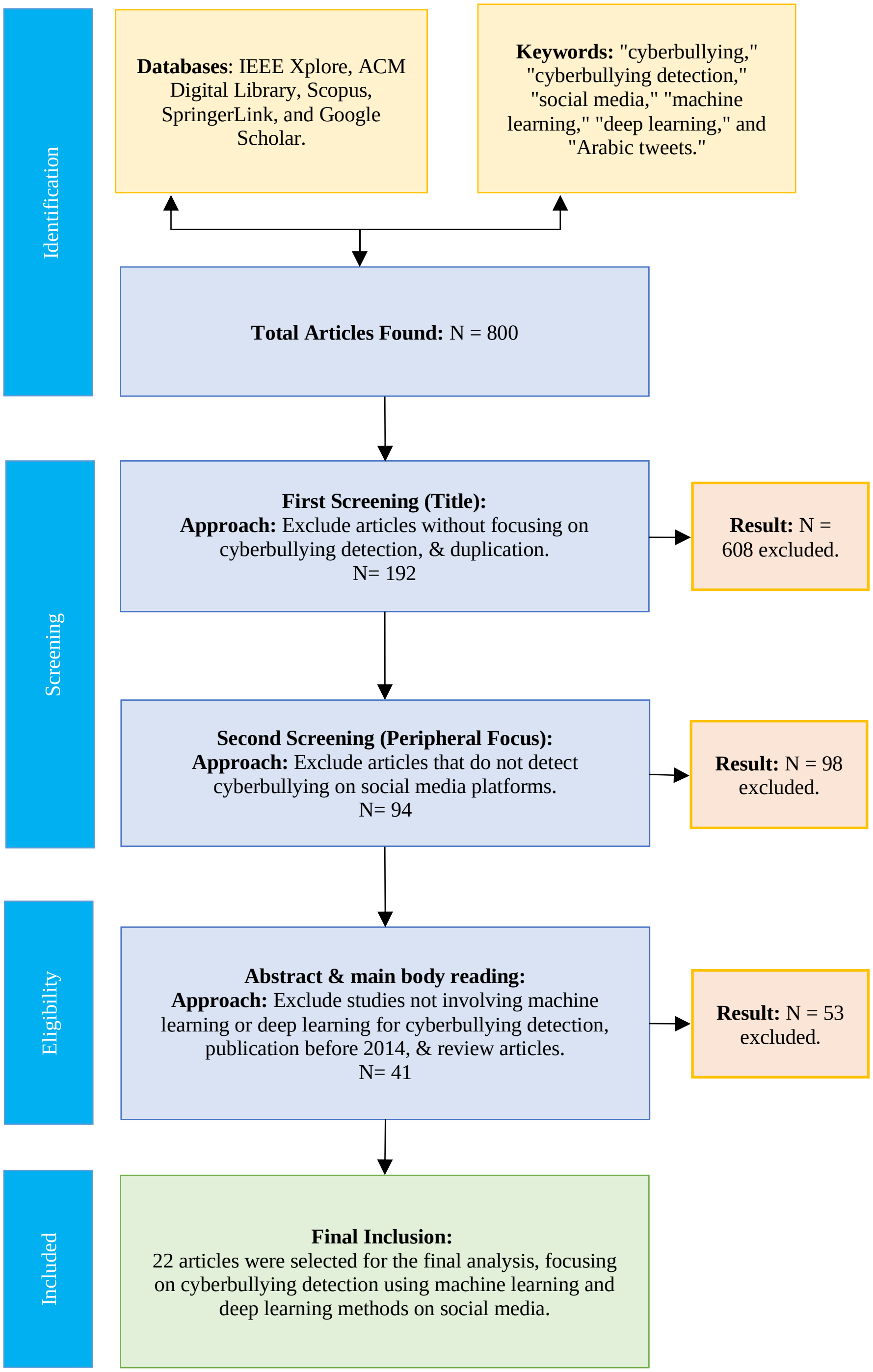
Identification
Databases: IEEE Xplore, ACM Digital Library, Scopus, SpringerLink, and Google Scholar.
Keywords: "cyberbullying," "cyberbullying detection," "social media," "machine learning," "deep learning," and "Arabic tweets."
Total Articles Found: N = 800
Screening
First Screening (Title):
Approach: Exclude articles without focusing on cyberbullying detection, & duplication.
N= 192
Result: N = 608 excluded.
Second Screening (Peripheral Focus):
Approach: Exclude articles that do not detect cyberbullying on social media platforms.
N= 94
Result: N = 98 excluded.
Eligibility
Abstract & main body reading:
Approach: Exclude studies not involving machine learning or deep learning for cyberbullying detection, publication before 2014, & review articles.
N= 41
Result: N = 53 excluded.
Included
Final Inclusion:
22 articles were selected for the final analysis, focusing on cyberbullying detection using machine learning and deep learning methods on social media.

Fig. 3 Literature review method

The review process followed these stages:

- Topical Relevance and Duplication: In the first screening, we excluded articles without a primary focus on cyberbullying detection and removed duplicates. This reduced the initial pool of 800 articles to 192.
- Focus on Cyberbullying: In the second screening, we excluded articles that did not specifically address cyberbullying detection on social media platforms. This further reduced the number of articles by 94, leaving 98.
- Methodological Rigor and Recency: During the eligibility stage, we conducted a detailed reading of abstracts and the main body of the remaining articles. Studies that did not involve machine learning or deep learning methods, those published before 2014, or those classified as review articles were excluded. This resulted in 53 exclusions, leaving a final pool of 41 articles.

Through this methodical approach, 22 articles were selected for detailed analysis, focusing on detecting cyberbullying using machine learning and deep learning methods, particularly on social media platforms.

8. Cyberbullying in Social Media: A Growing Concern

Social media platforms have been swiftly integrated into our daily lives because of their capacity to aid in communication, entertainment, and information dissemination. However, because of their easy access and anonymity, these platforms have also become hotspots for unpleasant behaviors, with cyberbullying being the most prominent [62]. Cyberbullying is a growing issue that affects people across all age groups but significantly affects teenagers. It is characterized by using digital communication tools to harass, intimidate, or harm others [63].

Research indicates that cyberbullying can be serious, causing victims to experience psychological discomfort, low self-esteem, and suicidal thoughts in extreme circumstances [64]. In addition, because social media can be accessed worldwide and is always active, victims frequently have the impression that there is no way to escape their tormentors. Therefore, it is of utmost importance to create efficient systems for recognizing and minimizing cyberbullying on various social media platforms [65].

As shown in [66], Twitter is among the most popular social media worldwide. In 2023, 550 million active users of Twitter across the world monthly. In addition to serving as a venue for exchanging information and views, it also serves as a platform for cyberbullying. A study indicates that Twitter has begun to become a platform for cyberbullying [20]. Cyberbullying on Twitter typically involves insulting comments, threats, or harassment, frequently targeting users' attributes or opinions [67]. Owing to Twitter's real-time and public characteristics, hurtful posts on Twitter have the chance to spread quickly and be viewed by a significant number of people. The harmful effects of cyberbullying are frequently exacerbated [20]. Therefore, developing machine-learning models that can detect and reduce cyberbullying on Twitter is essential to make the internet a more secure place for conducting online activities.

9. Literature Review

9.1. Machine Learning in English Text

The landscape of cyberbullying detection using English language texts has seen significant progress, with various Machine Learning (ML) and Deep Learning (DL) models being developed and applied in 2020 [68], introducing a machine-learning model aimed at identifying and curbing bullying on Twitter. This model utilized two classifiers, Naive Bayes and SVM, trained and tested on English-language social media bullying content. The SVM classifier outperformed the Naive Bayes model, achieving an accuracy of 71.25% in identifying true positives.

Around the same period, [69] examined how machine-learning techniques could be utilized for cyberbullying detection on Twitter. They used a wider range of classifiers, including Logistic Regression, Light Gradient Boosting Machine, Stochastic Gradient Descent, Random Forest, AdaBoost, Naive Bayes, and Support Vector Machine. The logistic regression classifier emerged as the most effective, achieving an F1 score of 0.928% and a median accuracy rate of approximately 90.57%.

[70] Developed a method based on sentiment analysis to identify cyberbullying on Twitter. Their study utilized Naive Bayes and Support Vector Machine approaches, operating on a dataset of tweets categorized as positive, negative, or neutral instances of cyberbullying. SVM classifiers surpassed NB classifiers in nearly all performance metrics across all language models. In particular, in the 4-gram language model, the SVM classifiers achieved an average accuracy of 92.02%, which was significantly higher than that achieved by the NB classifiers, 81.1%.

In 2021 [71], researchers devised a method employing natural language processing and machine learning algorithms to identify instances of cyberbullying within textual data. This study analyzed data from two forms of cyberbullying: derogatory comments directed at individuals on Wikipedia forums and offensive tweets containing hate speech on Twitter. In order to ascertain the optimal approach, they assessed four classifiers (Support Vector Machine, Logistic Regression, Random Forest, and Multilayer Perceptron). Moreover, they employed three feature extraction strategies (Bag of Words, Term Frequency-Inverse Document Frequency, and Word2Vec). The model achieves accuracy rates of over 90% for Twitter data and surpassing 80% for Wikipedia data.

9.2. Deep Learning in English Text

In the following year, 2021, they furthered their research [29] by introducing a method to detect stalking on Twitter, also using Sentiment Analysis. In this study, the approach incorporates Naive Bayes, Support Vector Machine, and Convolutional Neural Networks. The dataset used consisted of tweets labelled as positive, negative, or neutral instances of cyberbullying. The findings indicated that the CNN classifiers outperformed both the NB and SVM classifiers in almost all performance metrics across language models. Remarkably, in the 4-gram language model, the CNN classifiers achieved an average accuracy of 94.43%, compared with 81.1% and 91.64% achieved by the NB and SVM classifiers, respectively. This paper [5] introduced the CNN-CB algorithm. This unique method, built on a convolutional neural network, eliminates the need for feature engineering and asserts that similar words have analogous embeddings.

This resulted in more accurate cyberbullying detection, with experimental results showing a 95% improvement over traditional content-based methods.

This research [18] introduced a supervised machine-learning approach for identifying and addressing cyberbullying. Multiple classifiers were employed to train and identify bullying conduct. When the proposed method was tested on the cyberbullying dataset, it was found that the Neural Network surpassed the SVM, achieving an accuracy of 92.8%. Furthermore, the NN demonstrated superior performance compared to other classifiers that carried out similar tasks using the same dataset.

In 2021 [72], an automatic cyberbullying detection method based on a combined deep-learning model was proposed to identify aggressive behaviour. This novel approach leverages deep multichannel learning from three models: a bidirectional gated recurrent unit, a transformer block, and a convolutional neural network. The suggested approach demonstrated a remarkable precision rate of roughly 88% when tested on three popular hate-speech datasets.

In 2022 [28], both machine learning and deep learning were explored to automate the identification of cyberbullying comments. Model performance was measured using accuracy, precision, recall, and F1-score metrics. The study concluded that although SVM performed better when using machine learning methods, GRU marginally outperformed LSTM in the context of deep learning techniques. It was further confirmed that deep-learning techniques outpaced machine-learning techniques in terms of performance. The performance of the Gated Recurrent Units was particularly notable, with an accuracy of 95.47%.

In this study [73], eight novel emotional features were extracted, and a newly designed Deep Neural Network (DNN) with only three layers was employed to detect aggressive statements. The proposed DNN model was evaluated using the Cyber-Troll dataset. By integrating word embeddings with the eight emotional features, the model achieved substantial improvements in accuracy while maintaining a simplified and computationally efficient design. Compared to state-of-the-art models, the proposed model demonstrated an accuracy of 96%, outperforming existing approaches by a significant margin.

9.3. Machine Learning in Arabic Text

Machine Learning (ML) and Deep Learning (DL) models for identifying and preventing cyberbullying on Arabic-language social media platforms have significantly progressed in recent years. However, this is difficult to achieve. Notably, the lack of resources and information in Arabic compared to English makes it more challenging to create and improve these models [8].

Furthermore, owing to the complicated morphology and wide variety of dialects in Arabic, Arabic cyberbullying detection models are often less accurate than their English counterparts. This emphasizes the significance of allocating more resources to improve the efficacy of ML and DL tools for Arabic, guaranteeing equivalence with those created for English and enhancing the prevention of cyberbullying in Arabic-speaking communities.

The initial groundwork was laid in 2017 by [9], who developed a machine learning approach using a Support Vector Machine (SVM) and Naive Bayes (NB) algorithms to address cyberbullying on Arabic social media. This study pioneered the field using data collected from Facebook and Twitter. The experimental results were promising, indicating the feasibility of detecting Arabic cyberbullying through machine learning techniques, with SVM achieving a score of 0.934 and NB achieving a score of 0.901.

Furthermore,[10] employed predictive modelling to identify antisocial behaviour in Arabic YouTube comments. Utilizing a massive collection of offensive and non-offensive Arabic remarks, they trained their model using a Support Vector Machine classifier. This study added another layer to understanding Arabic-language cyberbullying, with the model attaining an accuracy of 90.05%.

In 2019, [11] presented an automatic machine-learning-based method for detecting cyberbullying in Arabic. Using the Naive Bayes classifier algorithm, they trained their model using real data collected from social media giants like Twitter and YouTube. The results were encouraging, demonstrating an accuracy of 0.959%, further validating the use of machine learning for Arabic cyberbullying detection.

In 2021 [12], supervised machine learning was used to develop a two-level classification model for violent Arabic text. The first level differentiated between violent and nonviolent content, whereas the second level classified violent text as cyberbullying or threatening. Using the SVM and NB algorithms, the authors experimented with various feature extraction techniques and stopword removal configurations. The results showed that SVM outperformed NB, thus solidifying it as a potent tool in this domain.

In 2023 [8], a machine-learning approach was proposed to identify cyberbullying on Arabic social media platforms. The researchers leveraged SVM and NB classifiers to detect instances of cyberbullying. They underscored the complexities of detecting cyberbullying in Arabic, owing to language intricacies and variability in user interactions. The SVM model was superior, with an accuracy rate of 95.742%.

9.4. Deep Learning in Arabic Text

Similarly, [74] proposed a deep-learning-based approach for identifying cyberbullying in Arabic using a Feed-Forward

Neural Network trained on an Arabic Twitter dataset. They reported improved results compared to previous studies and underlined the potential efficiency of deep learning methods for detecting cyberbullying in Arabic.

This research [75] utilized neural network models, specifically convolutional and recurrent neural networks, and pre-trained word embeddings for classifying cyberbullying instances within an Arabic news channel comments dataset. Their best models received an F1 score of 0.84% on a balanced dataset.

In 2024, [76] presents an integrated deep learning methodology that combines the most beneficial aspects of the fundamental models CNN, BLSTM, and GRU to effectively detect instances of cyberbullying. The proposed hybrid approach improves the accuracy across all evaluated datasets and can be integrated into various social media platforms to automatically identify cyberbullying cases within Arabic social datasets. It has the potential to substantially reduce cyberbullying incidents. The CNN-BLSTM-GRU model exhibited superior accuracy rates compared to other models utilizing DL and hybrid algorithms.

9.5. Machine and Deep Learning in Other Text

In 2017, [77] proposed a machine learning approach to identify instances of cyberbullying in Turkish social media posts. They used machine learning techniques, including Support Vector Machines (SVM), decision trees (C4.5), Naive Bayes Multinomial, and K-Nearest Neighbors (KNN), to analyze tweets and Instagram posts written in Turkish and identify instances of cyberbullying. They used information gain and chi-square feature selection techniques to boost classifier precision. They found that the cyberbullying identification accuracy increased when text words and emoticons were used as features. The Naive Bayes Multinomial classifier is the most effective classifier in terms of both accuracy and speed. The classification accuracy can be increased by as much as 84% using feature selection for a given dataset.

In 2019,[78] developed a Multimodal Cyberbullying Identification System that utilizes machine learning algorithms (MNB, LR, and SGD) to detect instances of cyberbullying in two different Indian languages: Hindi and Marathi. A prototype was developed with the datasets generated explicitly for these two languages. Utilizing this prototype, the researchers conducted experiments to detect cyberbullying instances in both languages. The test results revealed that Logistics Regression surpassed other algorithms in performance on these datasets, yielding an accuracy of up to 97% and an F1-score reaching 96% across multiple datasets for both languages.

This research [79] focuses on how specific functions of social media can be used to identify instances of cyberbullying. This study proposed using machine learning algorithms (SVM, LR, KNN, NBM, AdaBoost, and RF), a web annotation crowdsourcing program, and a data-gathering application to categorize cyberbullying content. After amassing a large dataset, they used a trained classifier to filter the unnecessary information. An online crowdsourcing platform was used to annotate data. The chi-square test examined the association between specific social media tools and cyberbullying. Classifier performance can be enhanced by incorporating social-media-related features into text-mining strategies. For example, the Support Vector Machine (SVM) accuracy can be improved by 3% when applied to datasets that include social media-related features.

In 2022 [80], machine learning and Natural Language Processing (NLP) methods were applied to identify instances of cyberbullying in comments made by Urdu Twitter users. They collected hateful tweets written in Urdu by Twitter users and used them to develop a dataset. The remarks were categorized into five distinct levels of offensiveness: innocuous (designated as 0), hostile/sexually abusive/general (designated as 1), disruptive in relation to blood, catastrophic incidents, mortality, and cruelty to animals (designated as 2), derogatory based on physical characteristics–body criticism/racial discrimination (designated as 3), and insulting in terms of political views (designated as 4). Features were extracted from characters and words using the N-gram method. Cyberbullying detection is achieved by applying several supervised machine-learning methods to the dataset (including XGBoost, Extra Tree Classifier, Multinomial NB, LR, Linear SVC, RF, K-NN, and Decision Tree). The results showed that the LR model performed better than the others, achieving 74.8% accuracy and 79.8% F1 score.

Using LSTM and GRU deep learning algorithms, this study analyzed Facebook comments in Bangla to identify cyberbullies [81]. After removing irrelevant comments, 7072 were retained in the analysis. The text was cleaned, punctuation was removed, tokenization was performed, stop words were eliminated, and stemming was performed as part of the data preprocessing steps. On the dataset, the GRU model outperformed the LSTM model with an accuracy of 83.55%. The literature review highlights the application of machine and deep learning to various forms of textual data, including English, Arabic, and other languages. A summary of the literature review is presented in (Table 1).

However, there are critical shortcomings in the literature. Among these is the incidence of imbalanced datasets, which can change the predictions made by models. Additionally, Arabic studies are less accurate in cyberbullying detection than English studies. Furthermore, there have been few studies on Arabic tweets, particularly in Saudi Arabia. Our examination of the literature indicates that no previous studies have examined the identification and evaluation of cyberbullying severity in Saudi Arabian tweets.

Table 1. The summary of literature review

Authors	Year	Methods	Dataset	Best Model	Accuracy
[68]	2020	Machine-learning model utilizing Naive Bayes and SVM classifiers	English	SVM	71.25%
[69]	2020	Various classifiers, including Logistic Regression, Light Gradient Boosting Machine, Stochastic Gradient Descent, Random Forest	English	LR	90.57%
[70]	2020	Sentiment analysis using Naive Bayes and Support Vector Machine approaches	English	SVM	92.02%
[29]	2021	Sentiment analysis using Naive Bayes, Support Vector Machine, and Convolutional Neural Networks	English	CNN	94.43%
[71]	2021	Machine learning SVM, RF, LR, MLR	English	SVM	90.20%
[5]	2018	CNN-CB algorithm based on convolutional neural networks	English	CNN-CB	95.00%
[18]	2019	Neural Network and SVM	English	NN	92.80%
[72]	2021	Combined deep learning model using bidirectional GRU, Transformer Block, and CNN	English	-	88.00%
[28]	2022	Comparison of machine learning and deep learning techniques	English	GRU	95.47%
[73]	2022	Deep neural network using word embedding and emotional features	English	DNN	96.00%
[9]	2017	SVM and NB algorithms for detecting Arabic cyberbullying	Arabic	SVM	93.04%
[10]	2018	Predictive modeling using SVM classifier for identifying antisocial behaviour in Arabic YouTube comments	Arabic	SVM	90.05%
[11]	2019	Naive Bayes classifier for detecting cyberbullying in Arabic	Arabic	NB	95.90%
[12]	2021	SVM and NB algorithms for two-level classification of violent Arabic text	Arabic	SVM	87.79%
[8]	2023	SVM and NB classifiers for identifying cyberbullying on Arabic social media platforms	Arabic	SVM	95.74%
[74]	2018	Deep learning approach using Feed-Forward Neural Network for identifying cyberbullying in Arabic	Arabic	FFNN	93.52%
[75]	2020	Convolutional and recurrent neural networks for classifying cyberbullying instances in an Arabic news channel comments dataset	Arabic	CNN-BLSTM-MAX	84.00%
[76]	2024	Several deep learning algorithms (LSTM, GRU, CNN–LSTM, CNN–BLSTM, LSTM–ATT, LSTM–TCN)	Arabic	CNN-BLSTM-GRU	85.00%
[77]	2017	Support Vector Machine (SVM), decision trees (C4.5), Naive Bayes Multinomial, and k Nearest Neighbours (KNN)	Turkish	NB	84.00%
[78]	2019	Machine learning (MNB, LR, and SGD)	Hindi and Marathi	LR	97.00%
[79]	2021	SVM, LR, KNN, NBM, AdaBoost, and RF	Turkish	SVM	90.10%
[80]	2022	XGBoost, Extra Tree Classifier, Multinomial NB, LR, Linear SVC, RF, K-NN, and Decision Tree	Urdu	LR	74.8%
[81]	2023	LSTM and GRU deep learning algorithms	Bangla	GRU	83.55%

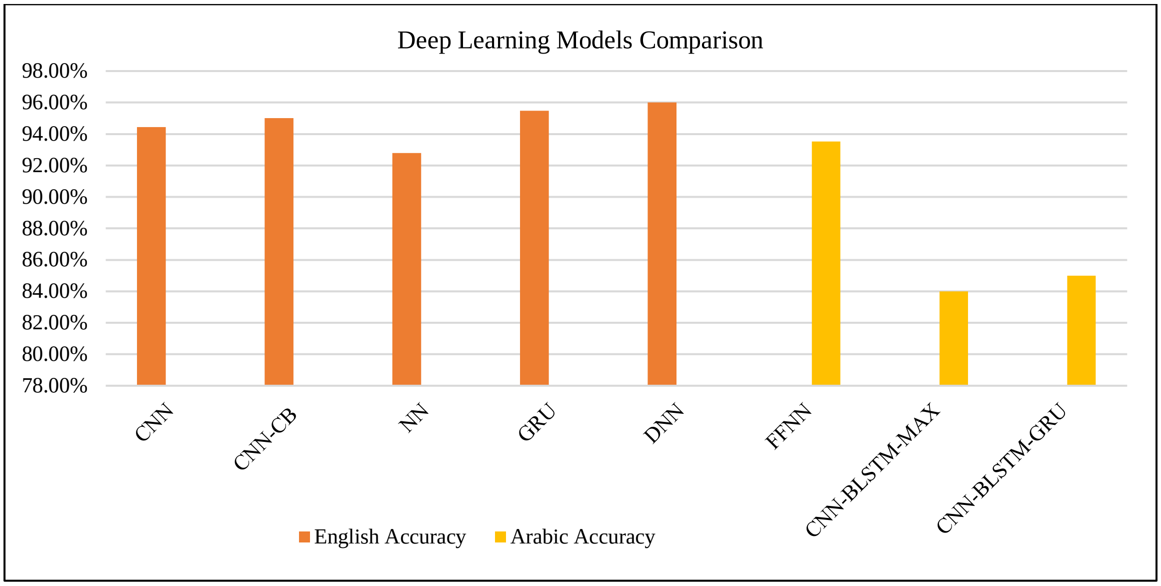

Fig. 4 Deep learning models comparison

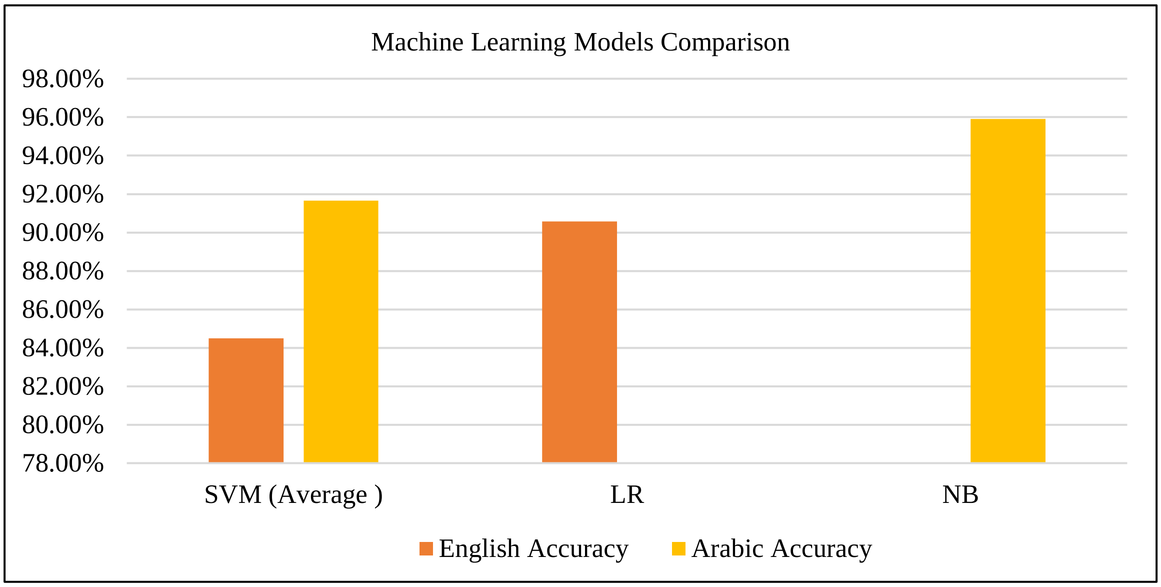

Fig. 5 Machine learning models comparison

10. Differences between Arabic and English Models in Cyberbullying Detection

In developing cyberbullying detection systems, the models for the Arabic and English languages exhibit distinct differences. These are influenced by each language's linguistic, cultural, and structural aspects-the disparities between the two present unique challenges and motivations for developing models.

Arabic is a morphologically rich language with a complex structure, making developing language models for cyberbullying detection more challenging. In contrast, English is comparatively straightforward and less intricate, as [39] noted. Additionally, the resources available for model training differ significantly between the two-English benefits from abundant annotated datasets tailored for cyberbullying detection. However, Arabic needs more such resources, further complicating the development of models for this language [8].

The rapid expansion of internet and social media usage accentuates the urgency to develop language models for Arabic. This development is essential for detecting and mitigating cyberbullying on Arabic social media platforms.

Additionally, promoting linguistic diversity is vital in AI and NLP systems. Since Arabic is among the most widely spoken languages globally, fostering models for this language ensures inclusive and universally accessible AI systems.

However, a limitation evident from the literature review is the comparative scarcity of resources available for Arabic text. This deficiency poses a challenge, as more training data can prevent machine learning models from underperforming. Such models may fail to capture the rich nuances and the diverse dialectal variations inherent to the Arabic language. This omission could compromise the effectiveness of NLP applications. A more substantial reservoir of training data for Arabic would enhance these models' accuracy and reliability. It would ensure that these models capture the linguistic intricacies and dialectal distinctions that are the Arabic language's hallmarks.

This enrichment leads to more proficient NLP applications benefiting developers and Arabic-speaking users. Addressing this gap requires an intensified effort from researchers to amass and train on Arabic data and to further the creation of lexicons and libraries. Encouragingly, tools and libraries for Arabic texts, such as Farasa, CAMeL Tools, AraNLP, and MADAMIRA, have been developed. These resources are instrumental in understanding and analyzing Arabic texts, stimulating researchers to train data further and develop models for Arabic text.

Developing Natural Language Processing (NLP) models for English and Arabic texts can be complex. A (Table 2) delineates this process into distinct sections, starting with data collection and traversing through various pre-processing and modelling stages before culminating in a performance evaluation. Each phase is labelled in the context relevant to English and Arabic, capturing each language's unique attributes and needs. For English text, tokenization, lowercasing, punctuation removal, and stop-word removal are crucial. Advanced processes like Named Entity Recognition, Part of Speech Tagging, and Dependency Parsing are also included. On the other hand, while many of these stages apply to Arabic text, the language demands specialized techniques to address its complexities. These include managing Arabic diacritics, morphological segmentation, text normalization, and light stemming, as highlighted by [82]. Thus, this table is a guide for building successful NLP models, which are nuanced to cater to the intricacies of both English and Arabic.

Table 2. The difference between English and Arabic models

Steps to Create Models	**English Model**	**Arabic Model**
Data Collection	✓	✓
Data Cleaning	✓	✓
Tokenization	✓	✓
Stopword Removal	✓	✓
Stemming	✓	✓
Lemmatization	✓	✓
Removing Punctuation	✓	✓
Removing Numbers	✓	✓
Handling Emojis	✓	✓
Removing URLs	✓	✓
Part-of-speech tagging	✓	✓
Named Entity Recognition (NER)	✓	✓
Syntactic Parsing	✓	✓
Handling Arabic Diacritics	✗	✓
Morphological Segmentation	✗	✓
Arabic Text Normalization	✗	✓
Arabic Stop-word removal	✗	✓
Arabic Light Stemming	✗	✓
Arabic Morphological Analysis	✗	✓
Feature Extraction	✓	✓
Splitting Data (Train/Test/Validation)	✓	✓
Model Selection	✓	✓
Model Training	✓	✓
Model Evaluation	✓	✓

Table 2 illustrates the steps in creating a Natural Language Processing model for English and Arabic text. As shown in the table, specific steps for creating an Arabic Natural Language Processing model include Transliteration, handling Arabic diacritics, morphological segmentation, text normalization, Arabic Stop-word removal, light stemming, and morphological analysis. See below for more details:

- Diacritics in Arabic, or "tashkeel" or "harakat", are small symbols above or below letters to denote vowels, consonant doubling, and other linguistic aspects. These marks can be removed during text preprocessing [83] to simplify the text. For instance, the sentence " السَّلامُ عَلَيْكُم ورَحْمَةُ اللهِ وَبَرَكاتُهُ" (with diacritics) can be shortened to"السلام عليكم ورحمة الله" (without diacritics).
- Morphological Segmentation: Breaking down words into their smallest meaningful components, known as morphemes. A word in Arabic can have roots, patterns, prefixes, suffixes, or other components [84].For example, the Word: "يستخدمون" (they use), Morphemes: "ي" (prefix indicating third person plural), "استخدم" (root word "use"), "ون" (suffix indicating masculine plural).
- Arabic Text Normalisation: This procedure entails changing several versions of a word into a standard form, such as replacing all forms of "Alif" (آ , إ ,أ) with a simple

“ا”. For example, the word “أهلا” can be normalized to “اهلا”.

- Arabic Stop-word Removal: Stop words are regularly used with no meaningful meaning and are frequently eliminated during text preprocessing. In Arabic, examples include (”من“ ,”في“ ,”و“), and others. For instance, after removing the stop-word, the sentence “الولد والبنت في المنزل” (The boy and the girl are in the house) can be turned into “القط الكلب المنزل” (The boy The girl The house).
- Arabic Light Stemming: This technique reduces inflected or derived words to their root or stem forms. Simple techniques such as eliminating known prefixes or suffixes are frequently used in light stemming [84]. For instance, the word “كتبت” (I wrote) can be stemmed to “كتب” (write).
- The Arabic morphological analysis process involves examining and evaluating many linguistic components within words, such as the root, pattern, prefixes, suffixes, and other relevant structural elements [84]. For instance, the word “يكتبونها” can be broken down to identify “كتب” as the root.
  “--ي--ون” as the pattern (marking a present tense verb with a masculine, third-person plural subject), and “ها” as a suffix (representing a feminine, third-person singular object).

In conclusion, detecting cyberbullying in Arabic poses several challenges, but the potential benefits of addressing this issue are substantial. The creation of robust models for identifying cyberbullying in Arabic could potentially foster a more secure online space, encourage respectful and empathetic interactions, and mitigate the detrimental effects of cyberbullying on psychological well-being.

11. Cyberbullying Detection Approach

Identifying cyberbullying in Arabic presents significant challenges due to the language's regional dialect variety and cultural differences, which introduce additional complexity to the detection process. However, recent research [8-12, 74-76] has significantly addressed these challenges by applying machine learning and deep learning techniques.

Researchers have analyzed various social media platforms, including Twitter, YouTube, and Instagram, to enhance the understanding and detection of harmful language in Arabic. This section examines key findings from these studies and the specific methodologies employed to foster safer online environments for Arabic-speaking users (Figure 6).

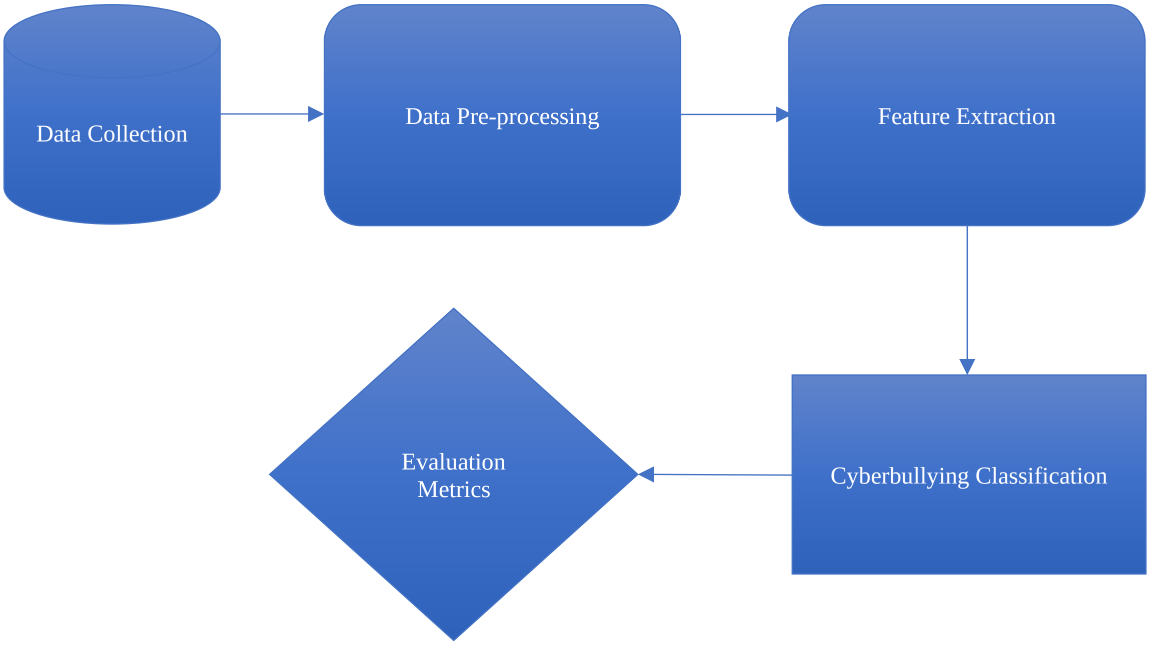

Fig. 6 Approach to cyberbullying detection

11.1. Data Collection and Preprocessing

The first step in detecting cyberbullying is data collection, which involves gathering content from various social media platforms, including Twitter, Facebook, and YouTube, to build a robust foundation for analysis (Figure 7). This initial dataset includes a variety of content types, providing a comprehensive basis for further analysis. However, raw data must undergo preprocessing to ensure consistency and relevance. This step involves organizing and standardizing the data through tokenization, normalization, manual labelling, stemming, segmentation, and removing irrelevant terms. These refinements optimize the dataset, enhancing the model's capacity to accurately detect key patterns by emphasizing significant linguistic features.

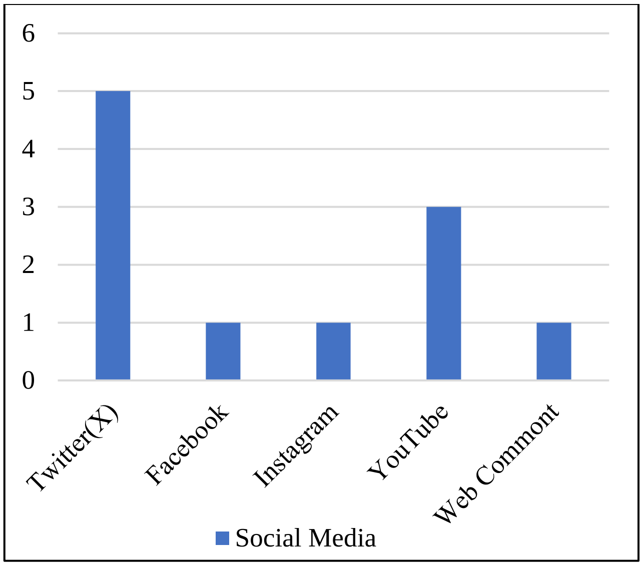

Fig. 7 Use of social media data in cyberbullying detection research

11.2. Feature Extraction Techniques

Following data collocation and pre-processing, the following essential step involves transforming the text into a format readable by a computer, known as feature extraction. Some researchers use basic techniques, like Term Frequency-Inverse Document Frequency (TF-IDF) or Bag of Words, which help models identify frequently occurring words and distinctive terms that may indicate specific patterns. Others opt for more advanced methods, such as word embeddings, which capture the contextual meanings of words-especially useful for languages where meanings can shift with dialect or phrasing. This process provides the model with a structured text representation, emphasizing essential terms and their relationships, ultimately enhancing pattern detection (Figure 8).

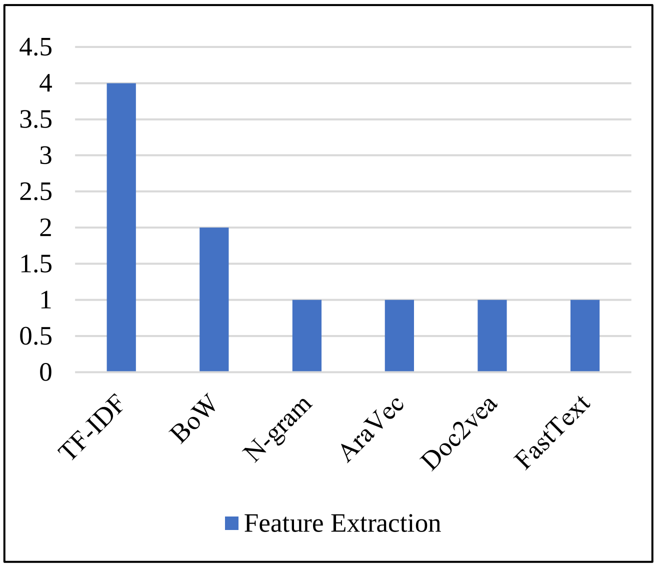

Fig. 8 Techniques for feature extraction in Arabic cyberbullying detection

11.3. Machine Learning and Deep Learning Models

In the detection phase, machine learning and deep learning techniques are applied to analyze text and identify patterns associated with specific language cues. Traditional machine learning methods, such as SVM and Naïve Bayes, are frequently used for their reliability in processing structured text data, effectively detecting common patterns and delivering consistent results across applications (Table 3**)**.

Conversely, deep learning models like CNNs and LSTMs are designed to capture more complex and subtle patterns, making them particularly valuable for interpreting nuanced language and contextual meanings within the text (Table 4). Both model types have shown strong performance in detecting and classifying relevant cues, offering complementary strengths that enhance the detection process.

Table 3. Overview of machine learning algorithms used in cyberbullying detection studies

Study	SVM	NB
[8]	✓	✓
[9]	✓	✓
[10]	✓	×
[11]	×	✓
[12]	✓	✓

Table 4. Overview of deep learning algorithms used in cyberbullying detection studies

Study	FFNN	CNN	LSTM	GRU	Hybrid Models
[74]	✓	×	×	×	×
[75]	×	✓	✓	✓	✓
[76]	×	✓	✓	✓	✓

11.4. Performance and Evaluation Metrics

Model performance is assessed using standard evaluation metrics, including accuracy, precision, recall, and F1-score (Table 5). These metrics evaluate each model's ability to identify relevant patterns while minimizing errors.

Accuracy indicates overall correctness, precision measures the model's effectiveness in avoiding false positives, recall assesses its sensitivity to detecting relevant instances, and the F1-score provides a balanced measure of precision and recall. Together, these metrics deliver a precise and reliable assessment of model effectiveness, helping to identify which approaches are most suitable for practical application.

Table 5. Overview of evaluation metrics in cyberbullying detection studies

Study	Accuracy	Precision	Recall	F1-Measure
[8]	✓	✓	✓	✓
[9]	×	✓	✓	✓
[10]	×	✓	✓	✓
[11]	✓	✓	✓	✓
[12]	✓	✓	✓	✓
[74]	✓	✓	✓	✓
[75]	✓	✓	✓	✓
[76]	✓	✓	✓	✓

12. Discussion

12.1. Key Insights from Arabic Cyberbullying Detection Studies

12.1.1. Data Sources

The research primarily relied on social media platforms like Twitter, YouTube, and Facebook, focusing on Arabic-speaking users. Some studies also incorporated content from alternative sources like Aljazeera.net comments and Instagram posts, offering a broad range of contexts for detecting offensive language.

12.1.2. Preprocessing Techniques

Common preprocessing steps included Normalization and Tokenization to address dialectal variations and the morphological complexity of Arabic. Stop-word removal and noise filtering to eliminate irrelevant elements, such as punctuation, URLs, diacritics, and emojis. Certain studies utilized Arabic-specific tools, including the Farasa toolkit and ARLSTem, to enhance accuracy in stemming and segmentation for improved language processing.

12.1.3. Feature Extraction

Methods varied based on the models employed. TF-IDF and Bag of Words (BoW) were frequently used for traditional machine learning models, helping to identify word importance. Word Embeddings (e.g., AraVec, FastText) and hybrid embeddings were prominent in studies using deep learning models, as they effectively capture semantic relationships between Arabic words.

12.1.4. Traditional Machine Learning Models

Support Vector Machine (SVM) and Naïve Bayes were commonly used for text classification, with SVM showing high precision and recall in detecting offensive content.

12.1.5. Deep Learning Models

Various deep learning architectures, including Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), and hybrid models (e.g., CNN-BLSTM-GRU) were explored to manage large datasets and enhance accuracy. The CNN-BLSTM-GRU hybrid model achieved the highest accuracy in multi-class classification, proving highly effective for the complexities of Arabic text structures.

12.1.6. Evaluation Metrics

Metrics such as accuracy, precision, recall, and F1 score were standard in evaluating model performance. Most studies reported high accuracy rates, with deep learning models excelling in nuanced classification tasks, reflecting their suitability for complex detection in Arabic cyberbullying contexts. For a detailed breakdown of the approaches and methodologies, see Table 6.

Table 6. Overview of ML and DL approaches in Arabic cyberbullying detection

Study	Data Sources	Preprocessing Techniques	Feature Extraction	Model Used	Evaluation Metrics
[9]	Twitter (4.93 GB), Facebook (0.98 GB)	Data Cleaning, Manual labelling, Normalizing	TF-IDF	SVM & NB	Precision, Recall and F-Measure
[10]	YouTube (15,050 comments)	Data Cleaning, Normalization, Tokenization	word-level features, N-gram features	SVM	Recall, Precision and F1- Score
[11]	Twitter and YouTube (25,000 comments)	Data Cleaning, Normalization, Stemming	Not specified	NB	Precision, Recall, F-Measure and Accuracy
[12]	Twitter (3,700 tweets)	Normalization, Noise Removal, Tokenization, Stop-word Removal	AraVec, TF-IDF	SVM & NB	Precision, Recall, F-Measure and Accuracy

[8]	Twitter and YouTube (30,000 comments)	Data Cleaning, Normalization Stemmed, Segmented	TF-IDF, BoW	SVM & NB	Precision, Recall, F1-score and Accuracy
[74]	Twitter (small data 4,913 and large data 34,890)	Not specified	Doc2Vec	Feed Forward Neural Network	Precision, Recall, F1-score and Accuracy
[75]	news channel Aljazeera.net (32,000 comments)	Data Cleaning, manual labelling, normalizing	Bag-Of-Words, Fasttext embeddings	CNNs, LSTMs, GRUs, hybrid CNN-RNN	Precision, Recall, F1-score and Accuracy
[76]	Instagram (46,000 comments)	Normalization, Data Cleaning, Tokenization, Stop-word Removal	Not specified	CNN-BLSTM-GRU	Precision, Recall, F1-score and Accuracy

12.2. Top Four High-Accuracy Studies in Arabic Cyberbullying Detection

Four of the eight studies reviewed on Arabic cyberbullying detection stood out for their solid methodologies and impressive results. Traditional models like Support Vector Machine (SVM) and Naïve Bayes (NB) showed reliable performance. For example, SVM [9] delivered balanced results with a precision of 0.934, recall of 0.941, and an F1-score of 0.927, proving effective for spotting subtle patterns in text. Similarly, NB [11] achieved the highest accuracy at 95.95%, showcasing its strength in handling probabilistic tasks.

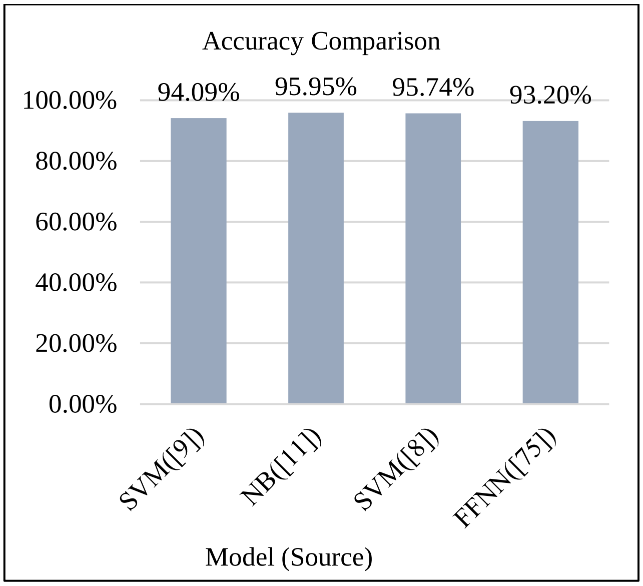

Fig. 9 Accuracy of top four arabic models

Another study [8] combined SVM with advanced preprocessing techniques, reaching an accuracy of 95.742%, though with slightly lower precision (0.92) and recall (0.84). On the deep learning side, a Feed Forward Neural Network (FFNN) [74] stood out, achieving the highest precision (0.959) and F1-score (0.956), demonstrating its ability to handle complex text structures. However, its accuracy was slightly lower at 93.2%. These results highlight the strengths of traditional and deep learning approaches, showing how important it is to tailor preprocessing and feature extraction to get the best outcomes in Arabic cyberbullying detection (Figures 9 and 10).

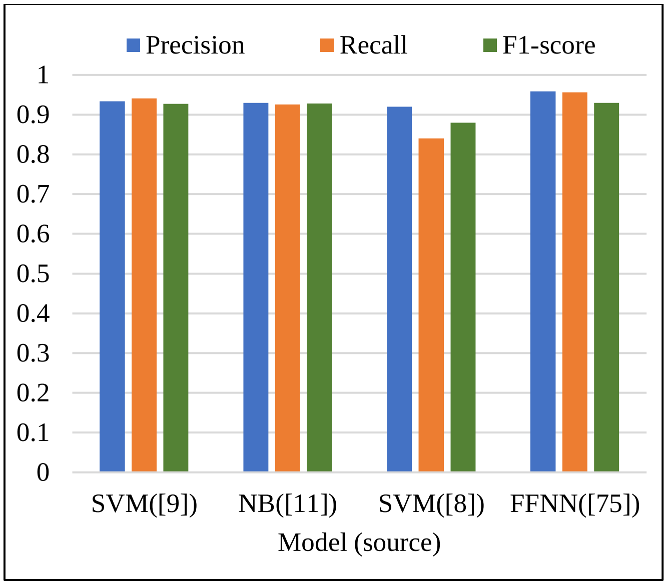

Fig. 10 Performance comparison of top four Arabic models

12.3. Limitation and Challenge

This literature review finds several significant limitations (Table 7). These include the frequency of imbalanced datasets, which can affect model predictions. The accuracy of cyberbullying detection in Arabic texts was lower than that in English texts. Furthermore, there is a notable lack of studies focusing primarily on Arabic tweets and, more specifically, Saudi Arabian tweets. According to our literature review, previous studies have not addressed the detection and severity assessment of cyberbullying in tweets from Saudi Arabia. In future work, we will aim to develop an automated module dedicated to detecting cyberbullying and determining its severity in Saudi Arabian tweets. Simultaneously, we will create a balanced dataset to improve the model's performance. There are several challenges to the development of this module. From **(**Table 7**)** referenced earlier, it is clear that the studies discussed in this literature review predominantly relied on imbalanced datasets, which could have influenced the accuracy and reliability of their models. A particular study by [75] emphasizes the impact of using an imbalanced dataset on model outcomes. Moreover, none of the existing studies has addressed the detection of cyberbullying severity.

Table 7. Limitations of literature review

Authors	Year	Limitation
[9]	2017	• Imbalance dataset. • The NB model detects 801 out of 2196 actual bullying. 31245 non-bullying out of 33077, which misses 1832. • The SVM detect 710 out of 2196 actual bullying and time-consuming 8 hours. 32479 non-bullying out of 33077, which misses 598. • No severity was detected.
[74]	2018	• Imbalance dataset: • For small dataset (1688 bullying, 3225 non-bullying), for large dataset (3015 as bullying, 31,875 as non-bullying) • No severity was detected. • No evaluation.
[10]	2018	• Imbalance data 39% offensive and 71% non- offensive. • No severity was detected.
[11]	2019	• Imbalance dataset. • No feature extraction. • No severity was detected.
[75]	2020	Used three versions of the data set: • First original: 26039 CB and 5653 non-CB (imbalance). • Second balance: 5653 non-CB and 5653 CB (which reduce cyberbullying to the same number of non-cyberbullying; the important bullying word will miss which effect the model's performer in future). • Third imbalance dataset (533 CB and 5653 non-CB). • No severity was detected.
[12]	2021	Imbalance dataset (used two levels to train data): • First level two classifies violent 1010 and not violent 990. • Second level three classifies 583 cyberbullying, 427 threatening, and 990 others. • No severity was detected.
[8]	2023	• Imbalance dataset 2279 cyberbullying out of 25221. • No severity was detected.
[76]	2024	• No severity was detected.

Recognizing severity is crucial for initiating appropriate responses; for example, when bullying is detected at a particularly severe level, the system might take measures, such as banning the offender and deleting the offending comments or tweets. Further insights into severity are discussed in the second challenge section. Given these observations, developing a model that identifies and gauges bullying severity is imperative. Using a balanced dataset is critical for obtaining accurate and reliable results.

The primary challenge lies in understanding and processing Arabic text. While Arabic is the mother tongue and dominant language in Saudi Arabia, it is inherently complex with its multiple dialects, writing systems, and morphological features. Using slang and informal speech on social media further complicates the situation. To address the intricacies of Arabic language processing, several tools and libraries have been developed (Table 8). While these resources can help overcome the challenge, processing the text may still be time-consuming due to the language's complexity. The table below lists some tools and libraries designed to aid in Arabic language processing.

Table 8. Tools and libraries used in Arabic text processing

NLP Tak	Example Tools/Libraries
Tokenization	NLTK, StanfordNLP
Handling Arabic Diacritics	Arabic NLP library (AraNLP)
Morphological Segmentation	Farasa, MADAMIRA, StanfordNLP, CAMel tool

Text Normalization	AraNLP
Light Stemming	NLTK (ISRI Arabic Stemmer), AraNLP, Farasa
Morphological Analysis	StanfordNLP, MADAMIRA, CAMel tool
Part-of-Speech Tagging	StanfordNLP, CAMel tool
Named Entity Recognition	StanfordNLP, CAMel tool
Sentiment Analysis	Deep Learning Libraries (Keras, PyTorch)

The tools and libraries previously mentioned process Arabic texts before being utilized in training machine learning models. Since the language used on social media might be a regional dialect within Saudi Arabia, processing can be time-consuming. Consequently, this step is the most intricate and requires significant time. It is essential to ensure the text is processed accurately to extract all the critical information necessary for training machine learning models to detect bullying in Arabic texts.

The second challenge is to assess the severity of cyberbullying. This presents a significant hurdle because of the need to understand the profound psychological impact of bullying behaviours. It is not only about identifying cyberbullying comments or threats; it requires a deeper understanding of the context, frequency, and power dynamics in which they manifest. An effective machine learning model must recognize these nuanced factors to gauge cyberbullying harm accurately. This understanding is pivotal for determining the severity of bullying and enabling machines to respond promptly. For instance, when bullying is detected at a high level, the machine should ban the individual and delete comments or tweets. If the severity is medium, the tweet should be deleted, and the person should be issued a warning. In cases of low-level bullying, the tweet should be straightforwardly deleted.

- A potential solution to this challenge is multilabel classification (Table 9). This approach categorizes each tweet with a severity score or label, such as High, Medium, Low, or Non-Cyberbullying. According to [85], tweets of a sexual /appearance are categorized as high severity, while racial /political tweets are designated as medium severity. Tweets targeting general are labelled as low severity.

Table 9. Categories of cyberbullying Severity

Category	Severity
Sexual /Appearance	High
Racial /Political	Medium
General	Low
Normal	Non-Cyberbullying

- The third challenge revolves around imbalanced datasets for cyberbullying detection. Our literature review highlighted a noticeable dataset imbalance during the data collection phase that persisted even after preprocessing. Specifically, instances of non-cyberbullying significantly outnumbered those of cyberbullying. This dataset imbalance can introduce bias into the training of a machine learning model and give inaccurate results, potentially undermining its performance in detecting less prevalent categories, such as cyberbullying.
- Data augmentation techniques or manual additions can be employed to address this challenge. These strategies generate synthetic data to balance the representations of the various classes. In the context of cyberbullying, this can create additional tweets that mimic bullying behaviours.

13. Problem Statement

As social media platforms have seen an exponential increase, cyberbullying has emerged as an extensive and harmful issue that detrimentally impacts individuals' mental health and overall well-being worldwide. The challenge of restraining this problem has escalated, considering the multitudes of freely accessible platforms, such as Twitter, teeming millions of users who can access these spaces anytime and anywhere. With 330 million active users each month, Twitter's popularity underscores the urgency of addressing cyberbullying on such an extensive scale. The ease of access to and use of this free social networking site allows cyberbullies to target victims with minimal resistance. Therefore, detecting and preventing cyberbullying is vital to ensuring a safer online environment for all users.

While substantial research has gone into identifying cyberbullying within English texts, the matter still needs more attention in Arabic texts. This research gap underlines the need for further exploration and development of detection methods specifically designed for Arabic, which could potentially help identify and address cyberbullying. The main hurdle in detecting cyberbullying in Arabic text on Twitter lies in the scarcity of resources, which includes a lack of publicly available data and insufficient training for machine learning models. This study aims to overcome these limitations by creating an efficient machine-learning-based method to identify cyberbullying in Arabic text on Twitter.

Our research focused on tackling the challenges of limited data and inadequate training to develop a reliable, accurate, and effective solution to combat cyberbullying. This study primarily focuses on Twitter-related cyberbullying incidents in Saudi Arabia.

14. Conclusion and Future Work

In conclusion, the review of the literature shows that there are three significant limitations in the field of cyberbullying detection within Arabic texts, especially in Saudi Arabian tweets: There are not many Arabic language resources; datasets are often not balanced; and there has been no previous work on measuring how severe cyberbullying is in cyberbullying detection methods. Addressing these challenges is crucial for advancing the efficacy and reliability of cyberbullying detection systems.

For future work, plan to develop a sophisticated model that detects cyberbullying in Arabic texts and categorizes incidents according to severity. This dual approach will involve the creation of a balanced dataset and applying advanced machine-learning techniques, which are essential for improving the model's accuracy and reducing bias. By incorporating severity detection, the model will provide more nuanced insights, enabling more appropriate and effective interventions at different levels of cyberbullying.

This research aims to fill critical gaps in the current understanding and capabilities of cyberbullying detection tools specifically tailored to the Arabic language. The successful implementation of this model has the potential to significantly enhance online safety and well-being for Arabic-speaking social media users, setting a new standard for cyberbullying detection technologies.

It also aims to raise awareness about the harmful effects of cyberbullying and inspire respectful and constructive online interactions within the Arabic digital community. The deployment of this automated cyberbullying detection system can act as a real-time solution for spotting harmful behaviors on Twitter. Ultimately, our goal is to help minimize the detrimental effects of cyberbullying on individuals and communities, cultivating a more inclusive and respectful online environment for Arabic speakers worldwide.

References

[1] Saudi Arabia Social Media Statistics, Global Media Insight, 2023. [Online]. Available: https://www.globalmediainsight.com/blog/saudi-arabia-social-media-statistics/#KSA_Social_Media_Statistics_2023_Top_Picks

[2] Ditch the Label, Cyberbullying Statistics: What They Tell Us, 2017. [Online]. Available: https://mx.ditchthelabel.org/cyber-bullying-statistics-what-they-tell-us

[3] Raju Kumar, and Aruna Bhat, "A Study of Machine Learning-Based Models for Detection, Control, and Mitigation of Cyberbullying in Online Social Media," *International Journal of Information Security*, vol. 21, no. 6, pp. 1409-1431, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[4] A.K. Jaithunbi et al., "Detecting Twitter Cyberbullying Using Machine Learning," *Annals of the Romanian Society for Cell Biology*, vol. 25, no. 4, pp. 16307-16315, 2021. [Google Scholar] [Publisher Link]

[5] Monirah Abdullah Al-Ajlan, and Mourad Ykhlef, "Deep Learning Algorithm for Cyberbullying Detection," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 9, pp. 199-205, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[6] Asif Ahmad Khan, and Aruna Bhat, "A Study on Automatic Detection of Cyberbullying Using Machine Learning," *6th International Conference on Intelligent Computing and Control Systems*, Madurai, India, pp. 1167-1174, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Vimala Balakrishnan, Shahzaib Khan, and Hamid R. Arabnia, "Improving Cyberbullying Detection Using Twitter Users' Psychological Features and Machine Learning," *Computers and Security*, vol. 90, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[8] Alanoud Mohammed Alduailaj, and Aymen Belghith, "Detecting Arabic Cyberbullying Tweets Using Machine Learning," *Machine Learning and Knowledge Extraction*, vol. 5, no. 1, pp. 29-42, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] Batoul Haidar, Maroun Chamoun, and Ahmed Serhrouchni, "Multilingual Cyberbullying Detection System: Detecting Cyberbullying in Arabic Content," *2017 1st Cyber Security in Networking Conference*, Rio de Janeiro, Brazil, pp. 1-8, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[10] Azalden Alakrot, Liam Murray, and Nikola S. Nikolov, "Towards Accurate Detection of Offensive Language in Online Communication in Arabic," *Procedia Computer Science*, vol. 142, pp. 315-320, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[11] Djedjiga Mouheb et al., "Detection of Arabic Cyberbullying on Social Networks using Machine Learning," *16th International Conference on Computer Systems and Applications*, Abu Dhabi, United Arab Emirates, pp. 1-5, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[12] Deema Alghamdi et al., "Automatic Detection of Cyberbullying and Threatening in Saudi Tweets using Machine Learning," *International Journal of Advanced and Applied Sciences*, vol. 8, no. 10, pp. 17-25, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[13] Pariwat Ongsulee, "Artificial Intelligence, Machine Learning and Deep Learning," *15th International Conference on ICT and Knowledge Engineering*, Bangkok, Thailand, pp. 1-6, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[14] Kamran Kowsari et al., "Text Classification Algorithms: A Survey," *Information*, vol. 10, no. 4, pp. 1-68, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[15] M. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text Classification using Machine Learning Techniques," *WSEAS Transactions on Computers*, vol. 4, no. 8, pp. 966-974, 2005 [Google Scholar] [Publisher Link]

[16] Charu C. Aggarwal, and ChengXiang Zhai, "A Survey of Text Classification Algorithms," *Mining Text Data*, Springer, Boston, USA, pp. 163-222, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[17] Cynthia Van Hee et al., “Automatic Detection of Cyberbullying in Social Media Text,” *PLOS One*, vol. 13, no. 10, pp. 1-22, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[18] John Hani et al., “Social Media Cyberbullying Detection Using Machine Learning,” *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 5, pp. 703-707, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[19] Taeho Jo, *Machine Learning Foundations Supervised, Unsupervised, and Advanced Learning*, Springer International Publishing, pp. 1-391, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[20] M.A. Al-Garadi, K.D. Varathan, and S.D. Ravana, “Cybercrime Detection in Online Communications: The Experimental Case of Cyberbullying Detection in the Twitter Network,” *Computers in Human Behavior*, vol. 63, pp. 433-443, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[21] Michele Di Capua, Emanuel Di Nardo, and Alfredo Petrosino, “Unsupervised Cyber Bullying Detection in Social Networks,” *IEEE 23rd International Conference on Pattern Recognition*, Cancun, Mexico, pp. 432-437, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[22] Xiaowei Gu, *A Self-Training Hierarchical Prototype-Based Approach for Semi-Supervised Classification*, Information Sciences, vol. 535, pp. 204-224, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[23] Vinita Nahar et al., “Semi-Supervised Learning for Cyberbullying Detection in Social Networks,” *Proceedings Databases Theory and Applications 25th Australasian Database Conference*, Australia, pp. 160-171, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[24] Yann Le Cun, Yoshua Bengio, and Geoffrey Hinton, “Deep Learning,” *Nature*, vol. 521, pp. 436-444, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[25] Shervin Minaee et al., “Deep Learning-Based Text Classification: A Comprehensive Review,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1-40, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[26] Celestine Iwendi et al., “Cyberbullying Detection Solutions Based on Deep Learning Architectures,” *Multimedia Systems*, vol. 29, pp. 1839-1852, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[27] Monirah A. Al-Ajlan, and Mourad Ykhlef, “Optimized Twitter Cyberbullying Detection Based on Deep Learning,” *IEEE 21st Saudi Computer Society National Computer Conference*, Riyadh, Saudi Arabia, pp. 1-5, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[28] K.G. Apoorva, and D. Uma, “Detection of Cyberbullying Using Machine Learning and Deep Learning Algorithms,” *IEEE 2nd Asian Conference on Innovation in Technology*, Ravet, India, pp. 1-7, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[29] Jalal Omer Atoum, “Cyberbullying Detection Neural Networks using Sentiment Analysis,” *IEEE International Conference on Computational Science and Computational Intelligence*, Las Vegas, NV, USA, pp. 158-164, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[30] Roman Egger, and Enes Gokce, *Natural Language Processing (NLP): An Introduction*, Applied Data Science in Tourism Interdisciplinary Approaches, Methodologies, and Applications, Springer, Cham, pp. 307-334, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[31] K.R. Chowdhary, *Natural Language Processing*, Fundamentals of Artificial Intelligence, Springer, New Delhi, pp. 603-649, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[32] Dipanjan Sarkar, *Text Analytics with Python, A Practitioner's Guide to Natural Language Processing*, Apress Berkeley, CA, pp. 1-674, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[33] Elizabeth D. Liddy, *Natural Language Processing*, In Encyclopedia of Library and Information Science, 2nd ed., Marcel Decker, NY, pp. 1-16, 2001. [Publisher Link]

[34] Yue Kang et al., “Natural Language Processing (NLP) in Management Research: A Literature Review,” *Journal of Management Analytics*, vol. 7, no. 2, pp. 139-172, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[35] Muhammad Abdul-Mageed, and Mona Diab, “AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis,” *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, Istanbul, Turkey, pp. 3907-3914, 2012. [Google Scholar] [Publisher Link]

[36] Hossam S. Ibrahim, Sherif M. Abdou, and Mervat Gheith, “Sentiment Analysis for Modern Standard Arabic and Colloquial,” *arXiv*, pp. 95-109, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[37] Kenneth R. Beesley, “Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001,” *ACL Workshop on Arabic Language Processing: Status and Perspective*, vol. 1, pp. 1-8, 2001. [Google Scholar]

[38] Tim Buckwalter, “Issues in Arabic Orthography and Morphology Analysis,” *Proceedings of the Workshop on Computational Approaches to Arabic Script-Based Languages*, Geneva, Switzerland, pp. 31-34, 2004. [Google Scholar] [Publisher Link]

[39] Ali Farghaly, and Khaled Shaalan, “Arabic Natural Language Processing: Challenges and Solutions,” *ACM Transactions on Asian Language Information Processing*, vol. 8, no. 4, pp. 1-22, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[40] Khaled Shaalan, “Rule-Based Approach in Arabic Natural Language Processing,” *International Journal on Information and Communication Technologies*, vol. 3, no. 3, pp. 11-19, 2010. [Google Scholar] [Publisher Link]

[41] Mohamed Elmahdy et al., “Survey on Common Arabic Language Forms from a Speech Recognition Point of View,” *Proceeding of International conference on Acoustics (NAG-DAGA)*, pp. 63-66, 2009. [Google Scholar] [Publisher Link]

[42] Khaled Shaalam, and Nizar Y. Habash, “Introduction to Arabic Natural Language Processing (Synthesis Lectures on Human Language

Technologies)," *Machine Translation*, vol. 24, no. 3-4, pp. 285-289, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[43] Kareem Darwish et al., "A Panoramic Survey of Natural Language Processing in the Arab World," *Communications of the ACM*, vol. 64, no. 4, pp. 72-81, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[44] Mohamed Abd Elaziz et al., *Recent Advances in NLP: The Case of Arabic Language*, Springer Cham, pp. 1-209, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[45] Sourabh Parime, and Vaibhav Suri, "Cyberbullying Detection and Prevention: Data Mining and Psychological Perspective," *IEEE International Conference on Circuits, Power and Computing Technologies*, Nagercoil, India, pp. 1541-1547, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[46] Vikas S. Chavan, and S.S. Shylaja, "Machine Learning Approach for Detection of Cyber-Aggressive Comments by Peers on Social Media Network," *IEEE International Conference on Advances in Computing, Communications and Informatics*, Kochi, India, pp. 2354-2358, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[47] Marilyn Campbell, and Sheri Bauman, *Cyberbullying: Definition, Consequences, Prevalence*, Reducing Cyberbullying in Schools International Evidence-Based Best Practices, pp. 3-16, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[48] Jaana Juvonen, and Elisheva F. Gross, "Extending the School Grounds?-Bullying Experiences in Cyberspace," *Journal of School Health*, vol. 78, no. 9, pp. 496-505, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[49] Andreas König, Mario Gollwitzer, and Georges Steffgen, "Cyberbullying as An Act of Revenge?," *Journal of Psychologists and Counsellors in Schools*, vol. 20, no. 2, pp. 210-224, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[50] Peter K. Smith et al., "Cyberbullying: Its Nature and Impact in Secondary School Pupils," *The Journal of Child Psychology and Psychiatry*, vol. 49, no. 4, pp. 376-385, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[51] Amrita Mangaonkar, Allenoush Hayrapetian, and Rajeev Raje, "Collaborative Detection of Cyberbullying Behavior in Twitter Data," *IEEE International Conference on Electro Information Technology*, Dekalb, IL, USA, pp. 611-616, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[52] Batoul Haidar, Maroun Chamoun, and Fadi Yamout, "Cyberbullying Detection: A Survey on Multilingual Techniques," *IEEE European Modelling Symposium*, Pisa, Italy, pp. 165-171, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[53] Samaneh Nadali et al., "A Review of Cyberbullying Detection: An Overview," *IEEE 13th International Conference on Intelligent Systems Design and Applications*, Salangor, Malaysia, pp. 325-330, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[54] Walisa Romsaiyud et al., "Automated Cyberbullying Detection using Clustering Appearance Patterns," *IEEE 9th International Conference on Knowledge and Smart Technology*, Chonburi, Thailand, pp. 242-247, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[55] Norulzahrah Mohd Zainudin et al., "A Review on Cyberbullying in Malaysia from Digital Forensic Perspective," *IEEE International Conference on Information and Communication Technology*, Kuala Lumpur, Malaysia, pp. 246-250, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[56] Deborah Goebert et al., "The Impact of Cyberbullying on Substance Use and Mental Health in a Multiethnic Sample," *Maternal and Child Health Journal*, vol. 15, pp. 1282-1286, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[57] Tanya Beran, and Qing Li, "The Relationship Between Cyberbullying and School Bullying," *The Journal of Student Wellbeing*, vol. 1, no. 2, pp. 16-33, 2007. [CrossRef] [Google Scholar] [Publisher Link]

[58] Justin W. Patchin, and Sameer Hinduja, Summary of Our Cyberbullying Research (2007-2023), 2024. [Online]. Available: https://cyberbullying.org/summary-of-our-cyberbullying-research

[59] Ana M. Giménez Gualdo et al., "The Emotional Impact of Cyberbullying: Differences in Perceptions and Experiences as A Function of Role," *Computers and Education*, vol. 82, pp. 228-235, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[60] Victoria Brown, Elizabeth Clery, and Christopher Ferguson, *Estimating the Prevalence of Young People Absent from School Due to Bullying*, National Centre for Social Research, 2011. [Google Scholar] [Publisher Link]

[61] Zachary Munn et al., "Systematic Review or Scoping Review? Guidance for Authors When Choosing Between a Systematic or Scoping Review Approach," *BMC Medical Research Methodology*, vol. 18, no. 1, pp. 1-7, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[62] Rebecca Dredge, John F.M. Gleeson, and Xochitl de la Piedad Garcia, "Risk Factors Associated with Impact Severity of Cyberbullying Victimization: A Qualitative Study of Adolescent Online Social Networking," *Cyberpsychology, Behavior, and Social Networking*, vol. 17, no. 5, pp. 287-291, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[63] Robin M. Kowalski et al., "Bullying in the Digital Age: A Critical Review and Meta-Analysis of Cyberbullying Research among Youth," *Psychological Bulletin*, vol. 140, no. 4, pp. 1073-1137, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[64] Sameer Hinduja, and Justin W. Patchin, "Bullying, Cyberbullying, and Suicide," *Archives of Suicide Research*, vol. 14, no. 3, pp. 206-221, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[65] Justin W. Patchin, and Sameer Hinduja, "Measuring Cyberbullying: Implications for Research," *Aggression and Violent Behavior*, vol. 23, pp. 69-74, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[66] Brian Dean, X (Twitter) Statistics: How Many People Use X?, 2025. [Online]. Available: https://backlinko.com/twitter-users#twitter-daily-active-users

[67] Homa Hosseinmard et al., “Detection of Cyberbullying Incidents on the Instagram Social Network,” *arXiv*, pp. 1-9, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[68] Rahul Ramesh Dalvi, Sudhanshu Baliram Chavan, and Aparna Halbe, “Detecting a Twitter Cyberbullying using Machine Learning,” *IEEE 4th International Conference on Intelligent Computing and Control Systems*, Madurai, India, pp. 297-301, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[69] Amgad Muneer, and Suliman Mohamed Fati, “A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter,” *Future Internet*, vol. 12, no. 11, pp. 1-20, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[70] Jalal Omer Atoum, “Cyberbullying Detection through Sentiment Analysis,” *IEEE International Conference on Computational Science and Computational Intelligence*, Las Vegas, USA, pp. 292-297, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[71] Varun Jain et al., “Detection of Cyberbullying on Social Media Using Machine Learning,” *IEEE 5th International Conference on Computing Methodologies and Communication*, Erode, India, pp. 1091-1096, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[72] Munif Alotaibi, Bandar Alotaibi, and Abdul Razaque, “A Multichannel Deep Learning Framework for Cyberbullying Detection on Social Media,” *Electronics*, vol. 10, no. 21, pp. 1-14, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[73] Umair Khan et al., “Aggression Detection in Social Media from Textual Data Using Deep Learning Models,” *Applied Sciences*, vol. 12, no. 10, pp. 1-16, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[74] Batoul Haidar, Maroun Chamoun, and Ahmed Serhrouchni, “Arabic Cyberbullying Detection: Using Deep Learning,” *IEEE 7th International Conference on Computer and Communication Engineering*, Kuala Lumpur, Malaysia, pp. 284-289, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[75] Benaissa Azzeddine Rachid, Harbaoui Azza, and Hajjami Henda Ben Ghezala, “Classification of Cyberbullying Text in Arabic,” *IEEE International Joint Conference on Neural Networks*, Glasgow, UK, pp. 1-7, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[76] Reem Albayari, Sherief Abdallah, and Khaled Shaalan, “Cyberbullying Detection Model for Arabic Text Using Deep Learning,” *Journal of Information and Knowledge Management*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[77] Selma Ayşe Özel et al., “Detection of Cyberbullying on Social Media Messages in Turkish,” *IEEE International Conference on Computer Science and Engineering*, Antalya, Turkey, pp. 366-370, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[78] Rohit Pawar, and Rajeev R. Raje, “Multilingual Cyberbullying Detection System,” *IEEE International Conference on Electro Information Technology*, Brookings, SD, USA, pp. 40-44, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[79] Alican Bozyiğit, Semih Utku, and Efendi Nasibov, “Cyberbullying Detection: Utilizing Social Media Features,” *Expert Systems with Applications*, vol. 179, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[80] Sara Khan, and Amna Qureshi, “Cyberbullying Detection in Urdu Language Using Machine Learning,” *IEEE International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering*, Lahore, Pakistan, pp. 1-6, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[81] Srabon Bhowmik Shanto, Mohammed Jahirul Islam, and Md. Abdus Samad, “Cyberbullying Detection using Deep Learning Techniques on Bangla Facebook Comments,” *IEEE International Conference on Intelligent Systems, Advanced Computing and Communication*, Silchar, India, pp. 1-7, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[82] Nizar Y. Habash, *Introduction to Arabic Natural Language Processing*, Springer Cham, pp. 1-170, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[83] Abdulrahman Ahmed Alzand, and Rosziati Ibrahim, “Diacritics of Arabic Natural Language Processing (ANLP) and Its Quality Assessment,” *IEEE International Conference on Industrial Engineering and Operations Management*, Dubai, United Arab Emirates, pp. 1-5, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[84] Arafat Awajan, “Arabic Text Preprocessing for the Natural Language Processing Applications,” *Arab Gulf Journal of Scientific Research*, vol. 25, no. 4, pp. 179-189, 2007. [Google Scholar] [Publisher Link]

[85] Bandeh Ali Talpur, and Declan O’Sullivan, “Cyberbullying Severity Detection: A Machine Learning Approach,” *PLOS One*, vol. 15, no. 10, pp. 1-19, 2020. [CrossRef] [Google Scholar] [Publisher Link]